Wikipediaを用いたメジャー情報と類似性のあるマイナー情報の検索手法の提案

著者	服部 祐基, 灘本 明代
雑誌名	甲南大学紀要.知能情報学編
巻	5
号	1
ページ	79-92
発行年	2012-07-20
URL	http://doi.org/10.14990/00001354

論　文

Wikipediaを用いたメジャー情報と類似性のあるマイナー情報の検索手法の提案

服部祐基[a], 灘本明代[b]

[a] 甲南大学大学院 自然科学研究科 情報システム工学専攻
神戸市東灘区岡本８－９－１, 658-8501
[b] 甲南大学 知能情報学部 知能情報学科
神戸市東灘区岡本８－９－１, 658-8501

（受理日 2012 年 5 月 14 日）

概 要

現在，インターネットの普及により，インターネット上には様々な情報が大量に存在している．その結果，認知度や知名度が低い情報が見つけにくい傾向にある．そこで本研究では認知度や知名度が低く見つけにくい情報をマイナー情報とし，このマイナー情報を検索し提示する手法の提案を行う．本論文ではマイナー情報の検索の第一歩として，検索対象ページを Wikipedia に限定し，Wikipedia からマイナー情報を検索する手法の提案を行う．具体的には，ユーザの興味や関心のあるクエリを入力とし，それと類似し且つマイナーである情報を「関連性に基づく検索」手法と「たとえ表現に基づく検索」手法の２つの手法を用いて検索を行う．ユーザの興味や関心のあることから検索を行うことによってマイナー情報を見つけやすくし，ユーザの新たな知識や興味の発見が可能にすることを目的としている．

キーワード: マイナー情報，類似検索，Wikipedia，たとえ表現

1　はじめに

現在，インターネット上には大量かつ様々な情報が溢れている．検索エンジンを用いて情報検索を行う際，認知度，知名度の高い情報は比較的簡単に得ることができる．それに対し，認知度，知名度が低い情報は見つけにくい傾向にある．本論文ではこのような認知度，知名度の低い情報をマイナー情報と呼ぶ．この見つけにくいマイナー情報の中にも重要な情報が含まれていると考えられるが，現在の検索システムではマイナー情報を見つけることが非常に困難である．例えばスポーツについて考える．世界には多くのマイナーなスポーツと少数のメジャーなスポーツが存在している．野球やサッカーなどのメジャーなスポーツはメディアに取り上げられることが多く，知っている人も多い．これらメジャーなスポーツに関する情報はインターネット上に多くのウェブページが存在しているため，これらの情報を見つけることは容易である．それに対しセパタクローやバンディのようなマイナーなスポーツは，メディアにはほとんど取り上げられないため，知っている人も少ない．そのためマイ

ナーなスポーツに関するウェブページも少なくこれらの情報を見つけることが困難になっている．しかしこのようなマイナー情報の中にもユーザにとって興味のある情報が含まれていると考えられるが，ユーザにとってどのように探せば良いかわからないという問題がある．そこで本研究ではこのように見つけにくいマイナー情報を検索しユーザに提示する手法の提案を行う．具体的には，ユーザの興味や関心のあることを入力とし，それと類似し且つマイナーである情報の検索を行う．これによりユーザの興味や関心のあることを入力とすることにより，見つけにくいマイナー情報を見つけ易くできる．つまりは，知らなかった情報や気づいていない情報を取得することができ，ユーザの新たな知識や興味の発見が可能になる．本研究では，マイナー情報の検索のはじめの一歩として Wikipedia に限定し，Wikipedia からマイナー情報を見つけることを行う．マイナー情報検索の大まかな流れを以下に，システムのフローを図 1 に示す．

1. ユーザは興味や関心のあるクエリを入力する．
2. 「関連性に基づく検索」手法を用いて，ユーザの入力したクエリの記事と関連性の高い記事を Wikipedia から取得する．この時，関連性の高い記事はリンクの解析と類似度計算を行うことにより決定する．
3. 「たとえ表現に基づく検索」手法を用いて，ユーザの入力したクエリを用いて，例えられている Wikipedia の記事を取得する．
4. 2 と 3 から取得した記事に対して編集回数，編集人数が少ない記事をマイナー情報と仮定し，抽出する．これらをマイナー情報の候補とする．
5. 4 で取得したマイナー情報の候補から入力したクエリのカテゴリ，上位概念が一致している記事を抽出する．それらをマイナー情報としてユーザに提示する．

以下，2 章では関連研究を，3 章ではマイナー情報の検索を，4 章ではフィルタリング手法を，5 章ではプロトタイプシステムを，そして 6 章では評価実験について述べる．そして 7 章でまとめと今後の課題について述べる．

2 関連研究

本研究では，Wikipedia からマイナー情報を検索する手法を提案している．大島ら [1] は複数の文書をクエリとし，それらと類似しているが異なる文書の検索を提案している．我々の提案はクエリから，類似し且つ異なるものを検索する点においては似ているが，本研究では入力されたクエリの Wikipedia の記事を用いて，それと類似しているが異なり，かつマイナー情報を見つけてくる点においては異なる．リンク関係の解析により Web ページの重要度，関連度を測る手法が多く提案されている．最も有名な手法としては PageRank [2] が上げられる．また Wikipedia のリンク構造の解析の手法も提案されている．Milne ら [3] は Wikipedia のリンクを用いたベクトルモデルである WLVM という手法を提案している．これは Wikipedia の記事のコンテンツを用いるのではなく，Wikipedia のリンク構造を唯一用いているという特徴のある手法である．また Chernov ら [4] は Wikipedia のカテゴ

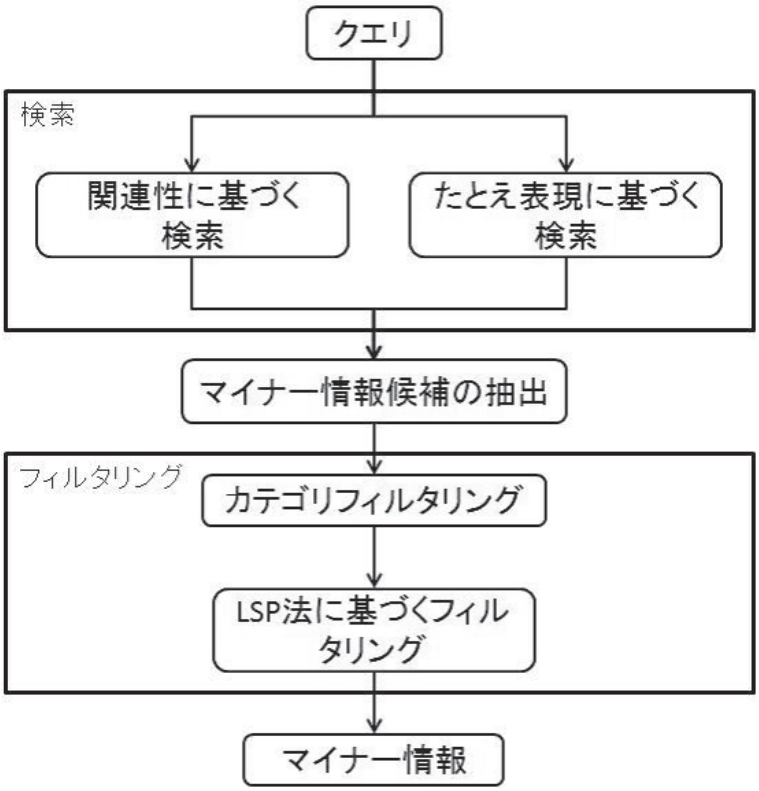

図 1: システムのフロー

リ間のリンクの解析を行うことにより Wikipedia から意味情報を抽出している．しかし我々の研究はマイナー情報を抽出することを行なっているため，これらの研究とは異なる．近年ユーザの嗜好に合わせて検索結果をリランキングする手法が多く提案されている．Zhuang ら [5] は検索クエリのログからクエリのコンテキストを作成することにより，任意のクエリに対して効果的に検索結果の順位を調整するための Q-ランクを提案している．Lee ら [6] はドキュメントのクラスタを使用して文書をリランキングする手法に基づく情報検索システムのモデルを提案している．これらのリランキング手法は下位にあるものを上位にランキングされる点でマイナー情報検索に似ているが，本研究では入力されたクエリと類似し且つマイナーな情報を検索する点で異なる．

3 マイナー情報の検索

本論文ではマイナー情報の検索として，Wikipedia からマイナー情報の抽出手法の提案を行う．マイナー情報を抽出する手法として,「関連性に基づく検索」と「たとえ表現に基づく検索」の２つの検索手法により入力したクエリと類似している記事を取得しする．そしてマイナー情報の候補の抽出方

法として，Wikipedia の記事の編集回数，編集人数を用いる．ここでは編集回数と編集人数が少ない記事をマイナー情報の候補として取得する．以下それぞれの手法について説明していく．

3.1 関連性に基づく検索

本節では入力されたクエリの記事と関連した記事を抽出することを行う．具体的には入力されたクエリの記事と関連性の高い記事を記事間のリンクグラフと類似度計算を用いて，以下の手順により関連のある記事を抽出する．

1. ユーザの入力したクエリの記事を基準としてリンクグラフを作成する．ここでユーザの入力したクエリの記事を基準ノードと呼ぶ．
2. 基準ノードにリンクしている記事や双方向にリンクしている記事（以下，インリンクノードと呼ぶ）は基準ノードに対して関連性を持っていると考えられる．そして基準ノードからのリンクのみの記事は，マイナーな記事である可能性が低いため，基準ノードのインリンクノード以外の記事を削除する．
3. インリンクノード内で，基準ノードのアンカー文字列が 1 回のみ出現する記事は関連が深くないと考え削除する．
4. 関連の高い記事同士は類似度が高くなると考えられるため，基準ノードとインリンクノードの記事間で類似度計算を行い，類似度がある閾値 α 以上のノードをマイナー情報の候補とする．本研究では以下のコサイン類似度を用いて，類似度 $\cos(x, y)$ を求める．

$$\cos(x, y) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}} \tag{1}$$

ここで，x をユーザが入力したクエリの記事，y をインリンクの関係にある記事，そして x_i は x における名詞 i の出現頻度，y_i は y における名詞 i の出現頻度である．ここで実験により閾値 α を 0.35 とし閾値 α 以上の記事を入力したクエリの記事と関連性の高い記事とし取得する．

図 2 ではノード C，E は基準ノードからのリンクのみのため削除する．ノード F は類似度が 0.2 であり閾値 0.35 以下であるため削除する．これにより図 2 の場合マイナー情報の候補となる記事はノード A, B, D となる．

3.2 たとえ表現に基づく検索

Wikipedia におけるマイナー情報は記事内の情報量がとても少ないものが含まれている．「関連性に基づく検索」手法では記事間の類似度計算により関連性を計っていたが，その際情報量が少ない記事と類似度計算をした場合，値が極端に少なくなるためマイナー情報であっても取得できない問題がある．そこで我々は Wikipedia 内のマイナー情報の記事の特徴として，あるマイナー情報を説明するために，クエリに似ている有名な情報に例えて表現している場合が多いことに注目する．この特徴を本

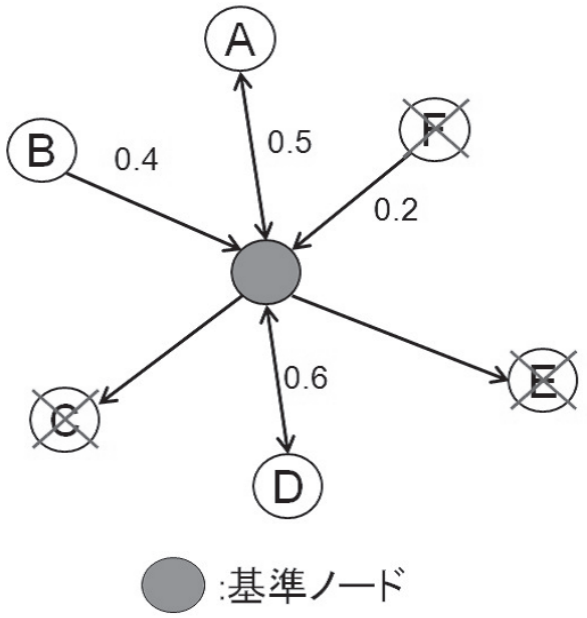

図 2: リンクグラフ

論文では「たとえ表現」と呼び，このたとえ表現を用いて，ユーザの入力したクエリからマイナー情報を抽出する．例えばスポーツにおいて「セパタクロー」というマイナースポーツの Wikipedia の記事では,「足のバレーボール」や「ルールはバレーボールと似て」というようにバレーボールに例えて表現がされている．また料理において「ムサカ」はギリシャ料理であるがあまり知られていない．「ムサカ」の Wikipedia の記事では,「グラタンに似た料理である」と表記されている．このようにたとえ表現を用いた文を読むだけでどのようなものであるかおおよそのイメージをすることができる．そこで「〜と似て」のようなたとえ表現を用いてマイナー情報の取得を行う．表 1 にたとえ表現の例と表記例を示す．たとえ表現を表す語には,「〜に類似」や「〜に似て」などの類似を表現している場合や,「〜の原型」,「〜の前身」,「〜の派生」などの起源を用いて表現している場合,「〜と…を組み合わせた」や「〜と…を融合した」などの複数のものを組み合わせて表現している場合などがある．そこでユーザが入力したクエリに対して，たとえ表現である表 1 の 7 パターンを拡張クエリとして付加し，検索を行う．例えば，入力が「バレーボール」であった場合には,「バレーボールに類似」や「バレーボールと似て」,「バレーボールの原型」のように拡張クエリを付加し，検索を行う．これにより得られた情報をマイナー情報の候補とする．

表 1: たとえ表現と表記例

たとえ表現の例	表記例（記事）
～に類似	サッカーに類似して（バンディ）
～と似て	バレーボールと似て（セパタクロー）
～と…を組み合わせた	クロスカントリーとライフル射撃を組み合わせた（バイアスロン）
～と…を融合	チェスとボクシングを融合したスポーツ（チェスボクシング）
～の原型	ゲートボールの原型（クロッケー）
～の前身	ボウリングの前身（ローンボウルズ）
～の派生	テニス等の派生（スカッシュ）

3.3 マイナー情報の候補の取得

ここでは「関連性に基づく検索」と「たとえ表現に基づく検索」で得られたクエリと類似した記事の中からマイナー情報の記事の取得を行う.「マイナー」の意味は大辞林では「[1] 規模や重要度が小さいさま．[2] あまり知られていないさま．有名ではないさま.」となっている．本研究ではあまり知られていない情報をマイナー情報と定義する．マイナー情報の記事を抽出するために，本研究では Wikipedia の編集回数と編集人数に着目する．メジャーな情報は知っている人も多く，Wikipedia の記事では編集する回数や人数が多くなるが，マイナーな情報は知っている人が少ないため Wikipedia の記事では編集する回数や人数が共に少なくなると仮定できる．この仮定を証明するために実験を行った．被験者 13 名に対し，ある 150 個の記事名を提示し，各記事名に対して以下 1～3 の 3 段評価をしてもらった．今回ユーザによる判定を容易にするために，スポーツデータを用いて実験を行なっている．

1 ： 知っている，だいたい知っている．

0 ： 名前は知っている．

-1 ： 知らない．

この実験の結果を表 2 に示す．ここでの平均とは，そのスポーツにおける評価の結果の合計を人数で割ったものである．よって “1” の値に近ければ知っている人が多くなり，“-1” に近ければ知っている人が少なくなる．よって平均がマイナスであればその記事はマイナー情報であると考えられる．

次にそれぞれ 150 個の記事に対して Wikipedia の記事の編集回数と編集人数を調べた．その結果を表 3 と図 3，図 4 に示す．表 3 の No1 から No4 はメジャー情報であるが，これらは編集回数・編集人数が多くなっている．それに対し No5 から No8 はマイナー情報であるが，編集回数・編集人数が少なくなっていることがわかる．また，図 3，図 4 では X 軸にそれぞれ編集回数と編集人数を，Y 軸に

表 2: 実験の結果

No	スポーツ名	平均	評価値 1 (人)	評価値 0 (人)	評価値 -1 (人)
1	フットサル	1	13	0	0
2	ハンドボール	1	13	0	0
3	水球	0.8462	11	2	0
4	ポロ	-0.6290	2	0	11
5	ロデオ	0.6154	8	5	0
6	カポエイラ	0.2308	5	6	2
7	フェンシング	1	13	0	0
8	バイアスロン	-0.3850	3	2	8
9	アーチェリー	1	13	0	0
10	ダーツ	1	13	0	0
11	ラクロス	0.8462	11	2	0
12	スカッシュ	0.7692	10	3	0
13	リュージュ	-0.5380	2	2	9
14	スノーカイト	-0.9230	0	1	12
15	カヌーポロ	-0.9230	0	1	12
：	：	：	：	：	：

は評価の平均としている．ここでは，編集回数，編集人数共に回数，人数が増えると，評価の平均が “1” である知っている人が多くなり，回数，人数が減ると共に，評価の平均が “-1” である知らない人が多くなっていることがわかった．これにより，メジャーな情報は編集回数と編集人数が多く，マイナーな情報は編集回数，編集人数が共に少なくなることが判明した．この結果より，たとえ表現に基づく検索と関連性に基づく検索により得られた記事の編集回数，編集人数を調べ，共に閾値 β 以下の記事はマイナー情報であるとする．ここで実験より閾値 β は編集回数を 180，編集人数を 80 とする．

4 フィルタリング

第 3 章から得られたマイナー情報の候補の記事の中には入力されたクエリと概念構造において同じクラスでないものが含まれている．ユーザに提示する情報として，入力されたクエリと概念クラスが同じものが好ましいと考えられる．そこでマイナー情報の候補からフィルタリングを行うことで入力したクエリと同じ概念クラスの記事を抽出することを提案する．ここで (1) カテゴリフィルタリング，(2) LSP 法に基づくフィルタリングの 2 つの手法を用いてマイナー情報のフィルタリングを行う．

表 3: 編集回数と編集人数

No	スポーツ名	編集回数	編集人数
1	野球	1087	507
2	サッカー	880	411
3	卓球	874	389
4	柔道	555	267
5	居合道	153	70
6	ペタンク	95	62
7	セパタクロー	76	46
8	ペサパッロ	28	24

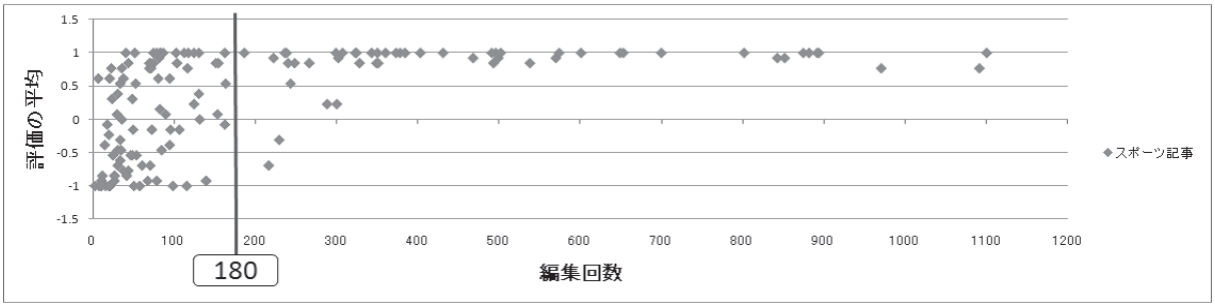

図 3: 編集回数と評価平均

4.1 カテゴリフィルタリング

Wikipedia の記事はすべてカテゴリに分類されている．我々はこのカテゴリに注目し，マイナー情報の候補の記事から入力されたクエリと概念クラスが同じであるマイナー情報を抽出する．入力されたクエリに対して Wikipedia でのカテゴリを取得する．ここで記事に対して上位 2 階層までのカテゴリをその記事の上位概念として取得する．そして第 3 章から得られたマイナー情報の候補のすべての記事に対して上位 2 階層までのカテゴリを取得し，それぞれの記事の上位概念として取得する．入力されたクエリのカテゴリとマイナー情報の候補の記事のカテゴリとの比較を行い，一致した記事をクエリと概念クラスが同じ記事として取得を行う．しかし記事により複数のクラスのカテゴリに含まれている場合がある．例えば「ボール」のカテゴリには「球技」が含まれており，クエリがバレーボールの場合，カテゴリには同じ「球技」が含まれ,「ボール」の記事を「バレーボール」の記事と同じ概念クラスと判断し取得してしまう．そこでこれらの記事を削除するために，LSP 法に基づくフィルタリング手法を用いる．

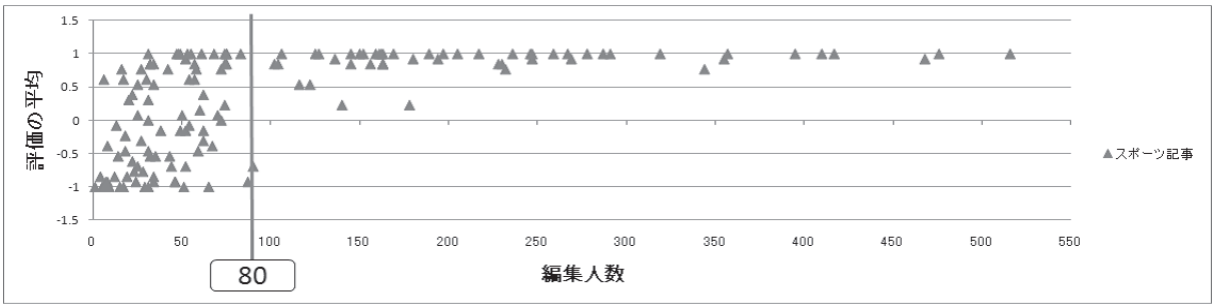

図 4: 編集人数と評価平均

4.2 LSP 法に基づくフィルタリング

Wikipedia のカテゴリを用いて取得された別の概念クラスの記事を削除するために，本研究では中山ら [7] の提案している LSP 法を用いる．中山らは Wikipedia 内の記事における文法に着目し，記事のリード部分を重要文と見なして解析する手法である LSP 法を提案している．Wikipedia 内の文章は多くの場合冒頭文が is-a 関係にある．この関係を利用し単語の上位概念を抽出することを行う．具体的には，記事の 1 文目の最後に出てきた名詞をその記事の上位概念とする．この手法を用いると例えば「ボール」の Wikipedia の記事の 1 文目は「ボール（ball、玉）はゲーム（球技や遊戯）などに使う球形の用具」となっており，最後に出てきた名詞である「用具」がボールの上位概念となる．クエリがバレーボールの場合,「用具」はバレーボールの上位概念にはならないため，異なる概念クラスであると判断できる．

以上２つのフィルタリング手法を用いて得られた記事をマイナー情報の記事としてユーザに提示する．

5 プロトタイプシステム

我々の提案手法を用いてプロトタイプシステムを作成した．プログラミング言語には Ruby を，ユーザインターフェースには CGI を用いた．システムの起動画面を図 5 に，クエリをバレーボールにした場合の出力例を図 6 に示す．まずユーザは興味のあるクエリを入力し，決定を押すとシステムはユーザの入力したクエリと似ているマイナー情報を検索する．そして得られたマイナー情報を一覧で表示する．またマイナー情報だけでなく，マイナー情報の Wikipedia の記事の概要を同時に表示している．ここでユーザがマイナー情報の一覧から一つの情報を選択するとその Wikipedia の記事が表示される．これによりそのマイナー情報についてより詳しく知ることができるようになる．

図 5: 起動画面

図 6: 出力画面

6 実験

提案手法の有用性を示すために，たとえ表現に基づく検索の評価実験とシステムの評価実験を行った．今回の実験では入力のクエリをスポーツとし，マイナースポーツを取得している．

6.1 たとえ表現に基づく検索の評価実験

たとえ表現に基づく検索で得られた記事がマイナー情報であるかを検証するために Wikipedia の編集回数と編集人数を用いて実験を行った．ここでは編集回数と編集人数を用いて，たとえ表現の評価を行う．3.3 節の実験によりマイナー情報と判定する編集回数と編集人数の閾値をそれぞれ 180，80 とし，たとえ表現に基づく検索から取得された記事のマイナーの度合いの評価を行う．入力となるクエリは任意の 7 種類のスポーツとしている．

結果と考察

たとえ表現に基づく検索より取得された記事名とその編集回数と編集人数との関係を表 4 に示す．結果よりたとえ表現に基づく検索から得られた記事の多くが編集回数，編集人数それぞれの閾値を下回る結果となった．たとえ表現に基づく検索より取得された記事のうち，フットサルやビーチバレー，ソフトボールなどのメジャーである記事もたとえ表現が用いられていることがわかった．しかしながら，フットサルとソフトボールは編集回数，編集人数ともに閾値以上となっている．そしてビーチバレーでも編集人数が閾値以上となっている．その為，編集回数，編集人数によりマイナースポーツの候補から削除されるため，提案手法は有用である事がわかった．

6.2 システムの評価実験

本節ではプロトタイプシステムを用いて，本システムの有用性を示す．本システムを評価する尺度として，再現率と適合率，F 値を用いる．再現率を求める際の正解データとして，表 2 における平均がマイナスの値を取っているマイナー情報の記事を用いている．ここで，マイナー情報の記事のみでは正解データが少ないため，これに合わせて人手で集めたマイナー情報 35 個を正解データとして加えている．実験データは，表 2 における平均が 1 に近い値を取っているメジャー情報の記事の中から 7 つのメジャー情報を入力データとして本提案手法を用いてマイナー情報検索を行った．

結果と考察

結果を表 5 に示す．表 5 より記事によって値に差が出ているが，再現率，適合率，F 値の平均は良い結果となった．しかしゴルフやラグビー，アーチェリーの再現率が他の記事に比べ低い値になっている．「LSP 法に基づくフィルタリング」において，記事の 1 文目の最後に出てきた名詞をその記事の上位概念としているが，1 文目が複文になっていた場合には正確に上位概念を取得することができないという問題がある．例えばサッカーのマイナースポーツの正解データである「フリースタイルフットボール」において，記事の 1 文目は「フリースタイルフットボール（Freestyle football）は、サッカーから派生したスポーツでサッカーボールを用いてパフォーマンスを行なうものである。」と複文で構成されている．この記事ではパフォーマンスが上位概念となりサッカーと同じ概念クラスと判断されずに取得できなかった．このため本来取得するべき記事を取得できずに再現率を下げる結果となった．このように LSP 法に基づくフィルタリングの問題点が明らかになった．

表 4: たとえ表現に基づく検索の取得例

入力	取得された記事	編集回数	編集人数
サッカー	バンディ	38	28
	フットサル	299	152
	サイクルサッカー	26	12
	電動車いすサッカー	33	21
	フリースタイルフットボール	5	3
バレーボール	セパタクロー	77	47
	インディアカ	43	25
	シッティングバレーボール	41	24
	ソフトバレーボール	20	14
	ビーチバレー	135	80
バスケットボール	ストリートバスケットボール	44	24
	ビーチバスケットボール	6	5
	ウォーターバスケットボール	6	6
	カヌーポロ	79	47
野球	クリケット	179	79
	スティックボール	6	6
	ソフトボール	354	94
テニス	タンブレロ	24	12
ラグビー	7 人制ラグビー	74	32
	タッチラグビー	16	5
柔道	サンボ	172	62
	クラッシュ	23	12

7 まとめと今後の課題

本論文では見つけにくいマイナー情報をユーザの興味や関心のあるものからマイナー情報を検索する手法の提案を行った．マイナー情報を検索し提示することを行うことで，ユーザの新たな知識の取得や新たな興味の発見が可能になる．今後の課題は以下の通りである．

- 他の Web ページへの応用
  本研究ではマイナー情報を抽出する対象として，Wikipedia に限定していたが，Web ページには Wikipedia の記事になっていないようなマイナー情報も多く含まれているため，今後は他の Web ページへの応用を考えていく必要がある．

表 5: 実験結果

スポーツ	再現率	適合率	F 値
サッカー	50%	71%	59%
ホッケー	57%	75%	65%
野球	67%	100%	80%
バレーボール	78%	88%	82%
テニス	57%	100%	73%
ゴルフ	40%	100%	57%
バスケットボール	100%	100%	100%
バドミントン	50%	100%	67%
ラグビー	17%	33%	22%
相撲	50%	75%	60%
フィギュアスケート	100%	100%	100%
アーチェリー	38%	100%	55%
平均	59%	87%	68%

- 出力結果の改善
  本システムでの出力では，マイナー情報のリストと Wikipedia の概要のみを提示しているが，マイナー情報の画像を一緒に載せることでその情報をよりイメージしやすかったのではないかと考えた．しかし画像検索によりマイナー情報を検索した結果，マイナーであるために画像が取得できないということがわかった．そのため Wikipedia の記事に画像が使われていた場合にはその画像も一緒に提示することも考えている．

- 評判情報の付加
  出力されたマイナー情報に対してより興味を持ってもらうために，評判情報を付け加えることを考えている． SNS のコミュニティを用いることで，そのマイナー情報のコミュニティ内での評判を抽出し，ユーザに提示する．これによりユーザがその情報に対してより興味や関心が増すと考えられる．

参考文献

[1] H. Ohshima, S. Oyama and K. Tanaka, “Sibling page search by page examples,” in *Proc. 9th International Conference on Asian Digital Libraries*, LNCS 4312, pp. 91-100, 2006.

[2] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” in *Proc. 7th International Conference on World Wide Web*, pp. 107-117, 1998.

[3] D. Milne, “Computing semantic relatedness using Wikipedia link structure,” in *Proc. New Zealand Computer Science Research Student Conference*, CD-ROM, 8 pages, 2007.

[4] S. Chernov, T. Iofciu, W. Nejdl and X. Zhuo, “Extracting semantic relationships between Wikipedia categories,” in *Proc. 1st International Workshop: SemWiki2006 - From Wiki to Semantics*, CD-ROM, 11 pages, 2006.

[5] Z. Zhuang and S. Cucerzan, “Re-ranking search results using query logs,” in *Proc. 15th International Conference on Information and Knowledge Management*, pp. 860-861, 2006.

[6] K. S. Lee, Y. C. Park and K. S. Choi, “Re-ranking model based on document clusters,” *Information Processing and Management*, vol. 37, no. 1, pp. 1-14, 2001.

[7] 中山浩太, 原隆治, 西尾章治郎, “自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジー自動構築,” 日本データベース学会論文誌, vol. 7, no. 1, pp. 67–72, 2008.